KU LEUVEN kulak

"The unstable ground below our feet, a possible solution"

TREE BASED FEATURE INDUCTION FOR BIOMEDICAL DATA

*Konstantinos Pliakos** *and Celine Vens*

Department of Public Health and Primary Care, KU Leuven Kulak

**konstantinos.pliakos@kuleuven.be*

Abstract

The performed study focuses on the issue of lacking variance in data representations used in biomedical data. The way it affects the machine learning scientific community is highlighted and a promising feature induction approach based on tree ensembles was proposed in order to handle that problem.

"Torture the data, and it will confess."

"Yes, but is it going to tell the truth?"

Problem statement

- Increase in the amount of the available research data.
- Some datasets suffer from lack in discrimination or even large amounts of duplicate feature vectors.
- There are genes or proteins for example, which despite having different functions, have exactly the same feature representation.

Data Study

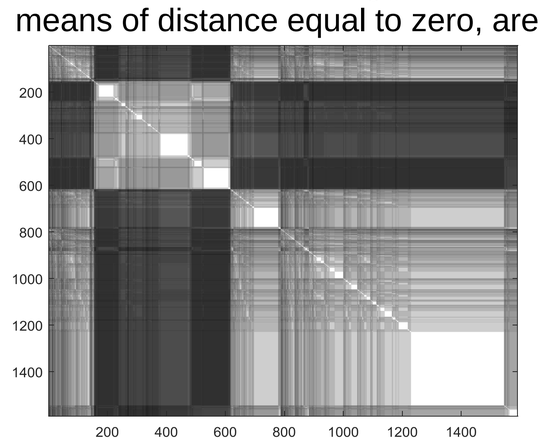
- In pheno dataset only 32.3% of the genes have unique representations.
- The most frequent feature vector appears 315 times, 197 in the training set and 118 in the test.

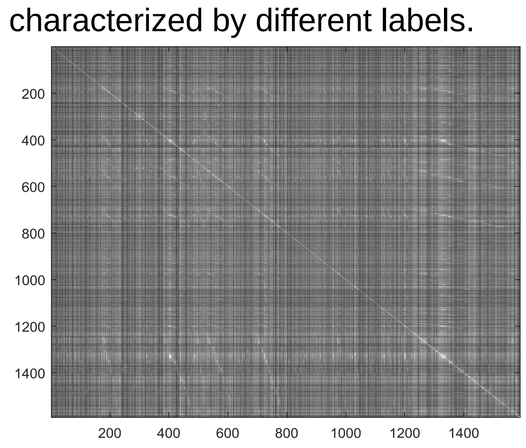
Context	Dataset	Nb of genes	Nb of unique gene representations
Gene function prediction (S. cerevisiae)	church	3755	**2352**
	pheno	1591	**514**
	hom	3854	3646
	seq	3919	3913
	struc	3838	3785
(A. thaliana)	scop	9843	9415
	struc	11763	11689
Drug protein interaction	drugs	1862	1779
	proteins	1554	**683**

***TABLE 1**. Datasets, the number of genes and their unique representations.*

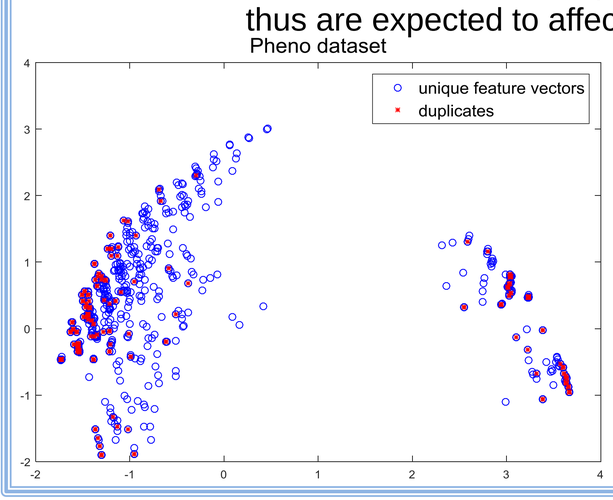
Data Visualization

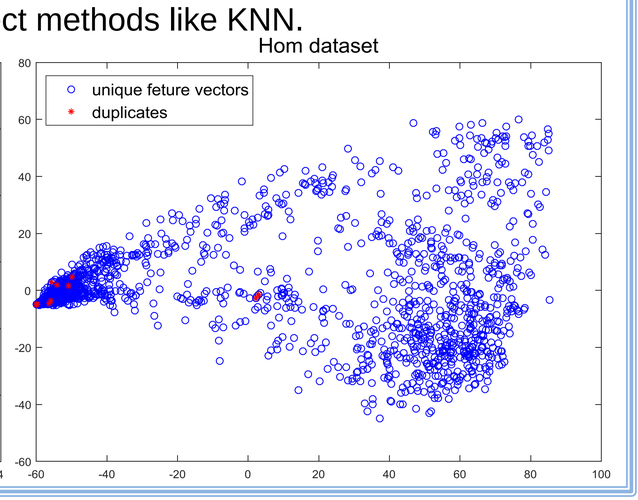
- Distance matrices for the features (left) and the labels (right) of the pheno dataset. It is shown that genes having the same feature representation, by means of distance equal to zero, are characterized by different labels.





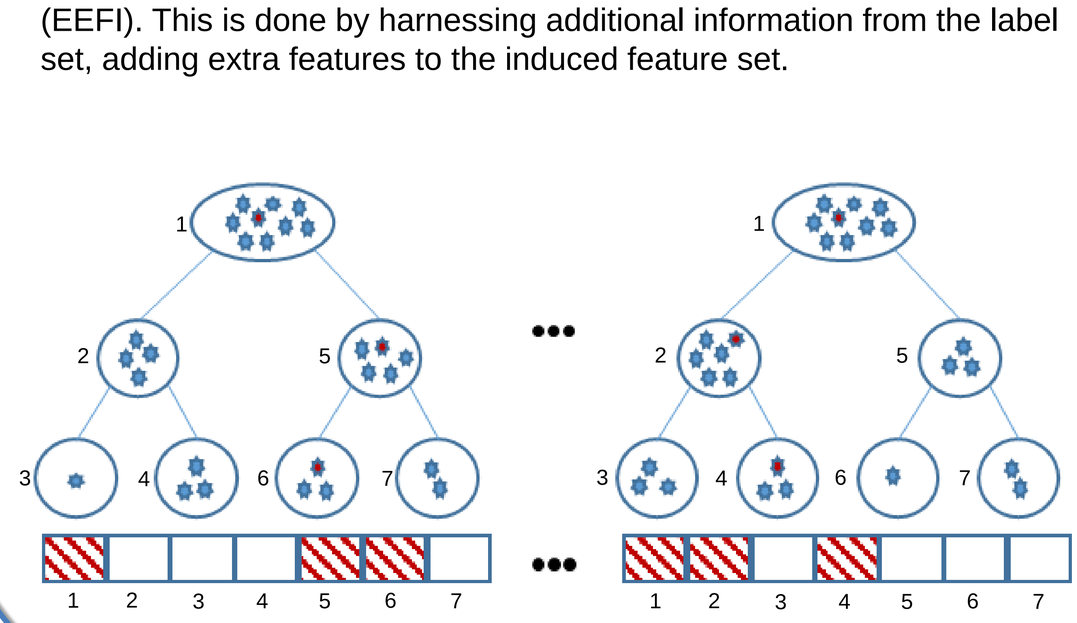
- Visualization of the data distribution using Multidimensional Scaling (MDS).
  - It is shown that these duplicates are similar to other genes, and thus are expected to affect methods like KNN.





Proposed Method

- We propose a method that generates a new feature set from an ensemble of Extremely Randomized Trees constructed over the data, coined as (EFI).
- The nodes of each decision tree of the ensemble are treated as clusters, containing all the samples that fall into that tree node. Next, binary feature vectors are generated, where each component represents the presence or absence of a sample in a cluster (node).
- Advantages of the method:
  - The new features are generated in an inductive manner.
  - They are label-aware.
  - Weights can be assigned to distill the information.
- The approach is further extended in order to tackle the issue of replicates (EEFI). This is done by harnessing additional information from the label set, adding extra features to the induced feature set.



Results

- Compared methods:
  - EFI (Extremely Randomized Feature Induction)
  - EEFI (Extended Extremely Randomized Feature Induction)
- A slight improvement is noted.

Dataset	original	EFI	EFI+PCA	EFI_joint
MN	0.27 / 0.83	0.33 / 0.83	0.32 / 0.82	0.30 / 0.84
PPI	0.21 / 0.84	0.21 / 0.84	0.19 / 0.82	0.22 / 0.84

Dataset	original	EEFI
pheno	0.16 / 0.83	0.17 / 0.85
DPI_1	0.21 / 0.84	0.21 / 0.84
DPI_2	0.03 / 0.60	0.03 / 0.61

***TABLE 2**. AUPRC / AUROC for the compared methods.*

Conclusion

The major points of our research:

- Inform the research community of this existing problem.
- Introduce a label informative feature induction approach based on tree ensembles.